A Large-Scale Multilingual Disambiguation of Glosses

José Camacho-Collados, Claudio Delli Bovi, Alessandro Raganato and Roberto Navigli

Department of Computer Science, Sapienza University of Rome
Via Regina Elena, 295 - 00161 Roma, Italy
`{collados,dellibovi,raganato,navigli}@di.uniroma1.it`

Abstract

Linking concepts and named entities to knowledge bases has become a crucial Natural Language Understanding task. In this respect, recent works have shown the key advantage of exploiting textual definitions in various Natural Language Processing applications. However, to date there are no reliable large-scale corpora of sense-annotated textual definitions available to the research community. In this paper we present a large-scale high-quality corpus of disambiguated glosses in multiple languages, comprising sense annotations of both concepts and named entities from a unified sense inventory. Our approach for the construction and disambiguation of the corpus builds upon the structure of a large multilingual semantic network and a state-of-the-art disambiguation system; first, we gather complementary information of equivalent definitions across different languages to provide context for disambiguation, and then we combine it with a semantic similarity-based refinement. As a result we obtain a multilingual corpus of textual definitions featuring over 38 million definitions in 263 languages, and we make it freely available at `http://lcl.uniroma1.it/disambiguated-glosses`. Experiments on Open Information Extraction and Sense Clustering show how two state-of-the-art approaches improve their performance by integrating our disambiguated corpus into their pipeline.

Keywords: Word Sense Disambiguation, Entity Linking, textual definitions, definitional knowledge, multilingual corpus

1. Introduction

In addition to lexicography, where their use is of paramount importance, textual definitions drawn from dictionaries or encyclopedias have been widely used in various Natural Language Processing tasks and applications. Some of the areas where the use of definitional knowledge has proved to be key in achieving state-of-the-art results are Word Sense Disambiguation (Lesk, 1986; Banerjee and Pedersen, 2002; Navigli and Velardi, 2005; Agirre and Soroa, 2009; Fernandez-Ordonez et al., 2012; Chen et al., 2014; Camacho-Collados et al., 2015b), Taxonomy and Ontology Learning (Velardi et al., 2013; Flati et al., 2014; Espinosa-Anke et al., 2016), Information Extraction (Richardson et al., 1998; Delli Bovi et al., 2015), Plagiarism Detection (Franco-Salvador et al., 2016), and Question Answering (Hill et al., 2015).

In fact, textual definitions (or glosses) are today widely to be found in resources of various kinds, from lexicons and dictionaries, such as WordNet (Miller et al., 1990) or Wiktionary, to encyclopedias and knowledge bases, such as Wikidata and OmegaWiki. These include Wikipedia itself: indeed, the first sentence of a Wikipedia article is generally regarded as the definition of its subject[1]. In any case, an accurate semantic analysis of a definition corpus is made difficult by the short and concise nature of definitional text. Furthermore, the majority of approaches making use of definitions are restricted to corpora where each concept or entity is associated with a single definition, while definitions coming from different resources are often complementary and might give different perspectives on the definiendum. Moreover, equivalent definitions of the same concept or entity may vary substantially according to the language, and be more precise or self-explanatory in some languages than others. This has the potential to be especially valuable in the context of disambiguation (Navigli, 2009), where highly ambiguous terms in one language may become less ambiguous (or even unambiguous) in other languages.

In this paper we bring together definitions coming from both different resources and different languages, and disambiguate them by exploiting their cross-lingual and cross-resource complementarities. Our goal is to obtain a large-scale high-quality corpus of sense-annotated textual definitions. In order to do this we leverage BabelNet[2] (Navigli and Ponzetto, 2012), a multilingual lexicalized semantic network obtained from the automatic integration of lexicographic and encyclopedic resources. Due to its wide coverage of both lexicographic and encyclopedic terms, BabelNet gives us a very large sense inventory for disambiguation, as well as a vast and comprehensive target corpus of textual definitions. In fact, as it is a merger of various different resources, BabelNet provides a large heterogeneous set of over 35 million definitions for over 250 languages from WordNet, Wikipedia, Wiktionary, Wikidata and OmegaWiki. To the best of our knowledge, this set constitutes the largest available corpus of definitional text.

We evaluate our sense-annotated corpus intrinsically, obtaining a disambiguation precision of over 90% on a random sample of definitions in three different languages, and extrinsically on Open Information Extraction and Sense Clustering tasks. Our experiments show the potential of exploiting our disambiguated glosses within the pipelines of two state-of-the-art systems, improving on their original performance.

2. Related Work

Among all resources using textual definitions, WordNet has definitely been the most popular and the most exploited

---

[1]According to the Wikipedia guidelines an article should begin with a short declarative sentence defining what (or who) the subject is and why it is notable.

[2]`http://babelnet.org`

to date. In fact, WordNet glosses have still been used successfully in recent work (Khan et al., 2013; Chen et al., 2015).

A first attempt to disambiguate WordNet glosses automatically was proposed as part of the eXtended WordNet project[3] (Novischi, 2002). However, this attempt's estimated coverage did not reach 6% of the total amount of sense-annotated instances. Moldovan and Novischi (2004) proposed an alternative disambiguation approach, specifically targeted at the WordNet sense inventory and based on a supervised model trained on the SemCor sense-annotated corpus (Miller et al., 1993). In general, the drawback of supervised models arises from the so-called *knowledge-acquisition bottleneck*, a problem that becomes particularly vexed when such models are applied to larger inventories, due to the vast amount of annotated data they normally require. Another disambiguation task focused on WordNet glosses was presented as part of the SensEval-3 workshop (Litkowski, 2004). However, the best reported system obtained precision and recall figures below 70%, which arguably is not enough to provide high-quality sense-annotated data for current state-of-the-art NLP systems.

In addition to annotation reliability, another issue that arises when producing a corpus of textual definitions is wide coverage. In fact, reliable corpora of sense-annotated definitions produced to date, such as the Princeton WordNet Gloss Corpus[4], have usually been obtained by relying on human annotators. The Princeton corpus of WordNet disambiguated glosses has already been shown to be successful as part of the pipeline in semantic similarity (Pilehvar et al., 2013), domain labeling (González et al., 2012) and Word Sense Disambiguation (Agirre and Soroa, 2009; Camacho-Collados et al., 2015b) systems. However, as new encyclopedic knowledge about the world is constantly being harvested, keeping up using only human annotation is becoming an increasingly expensive endeavor. With a view to tackling this problem, a great deal of research has recently focused on the automatic extraction of definitions from unstructured text (Navigli and Velardi, 2010; Benedictis et al., 2013; Espinosa-Anke and Saggion, 2014; Dalvi et al., 2015). On the other hand, the prominent role of collaborative resources (Hovy et al., 2013) has created a convenient development ground for NLP systems based on encyclopedic definitional knowledge. Nevertheless, extending the manual annotation of definitions to much larger and up-to-date knowledge repositories like BabelNet is not feasible. First of all, the number of items to disambiguate is massive; moreover, as the number of concepts and named entities increases, annotators would have to deal with the added difficulty of selecting context-appropriate synsets from an extremely large sense inventory. In fact, WordNet 3.0 comprises 117 659 synsets and a definition for each synset, while BabelNet 3.0 covers 13 801 844 synsets with a total of 40 328 194 definitions.

Instead, in this paper we propose an automatic disambiguation approach which leverages multilinguality and cross-resource information along with a state-of-the-art

multilingual Word Sense Disambiguation/Entity Linking system (Moro et al., 2014) and a vector-based semantic representation of concepts and entities (Camacho-Collados et al., 2015a). By exploiting these features, we are able to produce a large-scale high-quality corpus of glosses, automatically disambiguated with BabelNet synsets[5].

3. Methodology

The gist of our approach lies in the combination of different languages and resources for high-quality disambiguation. In fact, since many definitions are short and concise, the lack of meaningful context would negatively affect the performance of a Word Sense Disambiguation/Entity Linking system targeted at individual definitions.

To improve the data quality before the disambiguation step, we tokenize and Part-of-Speech (PoS) tag the definitions for a subset of languages:

Tokenization. We use the tokenization system available from the polyglot project[6] for 165 languages.

Part-of-Speech tagging. We train the Stanford tagger (Toutanova et al., 2003), for 30 languages using the available data from the Universal Dependencies project[7](Nivre, 2015).

Our disambiguation strategy is based on two steps: (1) all definitions are gathered together, grouped by definiendum and disambiguated using a multilingual disambiguation system (Section 3.1.); (2) the disambiguation output is then refined using semantic similarity (Section 3.2.).

3.1. Context-rich Disambiguation

As an example, consider the following definition of *castling* in chess as provided by WordNet:

Interchanging the positions of the king and a rook. (1)

The context in (1) is limited and it might not be obvious for an automatic disambiguation system that the concept being defined relates to *chess*: an alternative definition of *castling* where the game of *chess* is explicitly mentioned would definitely help the disambiguation process. Following this idea, given a BabelNet synset, we carry out a *context enrichment* procedure by collecting all the definitions of this synset in every available language and resource, and gathering them together into a single multilingual text.

We use a state-of-the-art graph-based approach to Entity Linking and Word Sense Disambiguation, Babelfy[8] (Moro et al., 2014), to disambiguate definitions after preprocessing and context-enrichment. Our methodology relies on the fact that, as shown in Section 3.1., disambiguation systems like Babelfy work better with richer context. When provided with the definition of Example (1) in isolation,

---

[3]http://www.hlt.utdallas.edu/~xwn/
[4]http://wordnet.princeton.edu/glosstag.shtml

[5]Note that BabelNet covers WordNet and Wikipedia among other resources, which makes our sense annotations expandable to any of these resources.
[6]http://polyglot.readthedocs.org/en/latest/Tokenization.html
[7]https://universaldependencies.github.io/docs/
[8]http://babelfy.org

	All languages	English	Spanish	French	Italian	Persian
Wikipedia	29 792 245	4 854 598	1 152 271	1 590 767	1 113 357	414 950
Wikidata	8 484 267	703 369	232 091	1 392 718	987 912	352 697
Wiktionary	281 756	281 756	-	-	-	-
OmegaWiki	115 828	29 863	22 446	12 777	14 763	11
WordNet	146 018	117 226	-	-	-	-
Total	**38 820 114**	**5 986 812**	**1 406 808**	**2 996 262**	**2 116 032**	**767 658**

Table 1: Number of disambiguated glosses by language (*columns*) and by resource (*rows*).

Babelfy incorrectly disambiguates *rook* as "rookie, inexperienced youth". However, by using additional definitions from other resources and languages, Babelfy exploits the added context and disambiguates *rook* with its correct *chess* sense. This approach is particularly advantageous for languages with low resources, where standard disambiguation techniques have not yet proved to be reliable, due to the shortage of annotated data.

3.2. Disambiguation Refinement

Babelfy outputs a set D of *disambiguated instances*, i.e. mappings from text fragments to items in the BabelNet sense inventory, each associated with a confidence score (*Babelfy score* henceforth). When Babelfy score goes below 0.7, a back-off strategy based on the *most common sense* is used by default for that instance. Our aim is to correct or discard these low-confidence instances using Semantic Similarity.

First, for each disambiguated instance $d \in D$ we compute a *coherence score* C_d. The coherence score is provided by Babelfy as the number of semantic connections from d to the rest of disambiguation instances in the semantic graph (normalized):

$$C_d = \frac{|\text{Disambiguated instances connected to } d|}{|\text{Disambiguated instances}| - 1} \quad (2)$$

We empirically set a coherence score threshold to 0.125 (i.e. one semantic connection out of eight disambiguated instances). Let L be the set of disambiguated instances below both Babelfy and coherence score thresholds (low confidence). In order to refine the disambiguated instances in L, we use NASARI[9] (Camacho-Collados et al., 2015a; Camacho-Collados et al., 2015b). NASARI provides vector representations for over four million BabelNet synsets built by exploiting the complementary knowledge of Wikipedia and WordNet. These semantic representations have proved capable of obtaining state-of-the-art results in various lexical semantics tasks such as Semantic Similarity, Sense Clustering and Word Sense Disambiguation. We consider those instances in L for which a NASARI vector can be retrieved (virtually all noun instances), and compute an additional score (*NASARI score*). First, we calculate the centroid μ of all the NASARI vectors for instances in $D \setminus L$. Then, for each disambiguated instance $l \in L$,

we retrieve all the candidate senses of its surface form in BabelNet and calculate a NASARI score N_s for each candidate sense. N_s is calculated as the cosine similarity between the centroid μ and its corresponding NASARI vector $NASARI(s)$:

$$N_s = Sim(\mu, NASARI(s)) \quad (3)$$

The NASARI score allows us to both discard low-confidence disambiguated instances and correct the original disambiguation output by Babelfy in some cases. Then, each $l \in L$ is re-tagged with the sense obtaining the highest NASARI score:

$$\hat{s} = \underset{s \in S_l}{\operatorname{argmax}} N_l \quad (4)$$

where S_l is the set containing all the candidate senses for l. For what concerns the high-precision disambiguated glosses release (see Section 6.) we set the NASARI threshold to 0.75. Considering example (1) again, Babelfy does not provide a high-confidence disambiguation for the word *king*, which is then incorrectly disambiguated using the most common sense strategy. However, the error is fixed during the refinement step: our system accurately selects the *chess* sense of *king* thanks to its high semantic connection with the disambiguated instances in $D \setminus L$.

4. Statistics

The output of our disambiguation procedure is a corpus of 38 820 114 glosses extracted from BabelNet (corresponding to 8 665 300 BabelNet synsets), covering 263 languages and 5 different resources (Wiktionary, WordNet[10], Wikidata, Wikipedia[11] and OmegaWiki) and including 249 544 708 annotations from the BabelNet sense inventory (6.4 annotations per definition on average). Table 1 reports some general statistics of the complete corpus of disambiguated textual definitions and for five sample languages: English, Spanish, French, Italian and Persian.

The number of disambiguated instances, before and after the refinement step, are displayed in Tables 2 and 3, organized, respectively, by language and Part-of-Speech (PoS). *Babelfy* and *NASARI* refer to the instances disambiguated by the two respective approaches and *MCS* to the

---

[9]We use the 2.1 release version of the *NASARI-embed* vectors, downloaded from http://lcl.uniroma1.it/nasari

[10]Including Open Multilingual WordNet.

[11]Definitions from Wikipedia include both first sentences of Wikipedia articles and definitions coming from Wikipedia's disambiguation pages.

		All languages	English	Spanish	French	Italian	Persian
Before refinement	Babelfy	174 256 335	39 096 127	9 006 888	11 178 328	8 892 763	3 766 754
	MCS	75 288 373	19 724 340	5 164 557	7 064 210	4 525 610	1 524 267
	Total	**249 544 708**	**58 820 467**	**14 171 445**	**18 242 538**	**13 418 373**	**5 291 021**
After refinement	Babelfy	144 637 032	33 260 600	7 029 173	8 735 298	7 106 414	3 085 804
	NASARI	18 392 099	4 680 745	1 353 494	1 865 920	1 301 370	330 917
	Total	**163 029 131**	**37 941 345**	**8 382 667**	**10 601 218**	**8 407 784**	**3 416 721**

Table 2: Number of annotations by language (*columns*) and by type (*rows*) before and after refinement.

		All content words	Nouns	Verbs	Adjectives	Adverbs
Before refinement	Babelfy	174 256 335	158 310 414	4 368 488	10 646 921	930 512
	MCS	75 288 373	56 231 910	8 344 930	9 256 497	1 455 036
	Total	**249 544 708**	**214 542 324**	**12 713 418**	**19 903 418**	**2 385 548**
After refinement	Babelfy	144 637 032	140 111 921	1 326 947	3 064 416	133 748
	NASARI	18 392 099	18 392 099	-	-	-
	Total	**163 029 131**	**158 504 020**	**1 326 947**	**3 064 416**	**133 748**

Table 3: Number of annotations by Part-of-Speech (PoS) tag (*columns*) and by type (*rows*) before and after refinement.

instances which were disambiguated using the Most Common Sense (MCS) heuristic. After refinement, 24.7% of the low-confidence noun annotations are fixed using semantic similarity (see Section 3.2.). Assuming the coverage of our first disambiguation step (see Section 3.1.) to be 100%[12], the coverage of our system after the refinement step is estimated to be 65.3%. As shown in Table 3, discarded annotations mostly include verbs, adjectives and adverbs, often harder to disambiguate as they are not directly related to the definiendum. In fact, the coverage of noun instances after refinement is estimated to be 73.9%.

5. Evaluation

5.1. Intrinsic evaluation

We first carry out an intrinsic evaluation of the resource, by manually assessing the quality of disambiguation on some randomly extracted samples of definitions. We rely on three human judges and evaluate samples of 100 items for three languages. We evaluated the disambiguation output before and after the refinement step, and compared against a baseline where each definition is disambiguated in isolation with Babelfy. Table 4 reports the evaluation on the three sample languages: English, Spanish and Italian. Although the disambiguation of context-free definitions improves only slightly with respect to the disambiguation of definitions in isolation, this improvement is consistent across languages. Furthermore, our system significantly increases the precision after the refinement step. Refinement reduces the coverage by 35% for English, and by 43% for Spanish and Italian, but increases precision by almost 11% for English, 20% for Spanish and 13% for Italian.

---

[12]There is no straightforward way to estimate the coverage of a disambiguation system automatically. In our first step using Babelfy, we provide disambiguated instances for all content words (including multi-word expressions) from BabelNet and also for overlapping mentions. Therefore the output of our first step, even if not perfectly accurate, may be considered to have full coverage in comparison with our refinement step.

	English		Spanish		Italian	
	Prec.	Cov.	Prec.	Cov.	Prec.	Cov.
Definitions in isolation	84.2	100	74.6	100	77.6	100
Context-rich defs. pre-refin.	84.3	100	74.7	100	78.0	100
Context-rich defs. post-refin.	**95.1**	64.8	**95.0**	57.3	**91.1**	56.8

Table 4: Disambiguation precision (Prec.) and coverage (Cov.) percentage (%) of the three different disambiguation strategies on the 300 sample definitions.

5.2. Extrinsic evaluation

The sense-annotated corpus of definitions is also evaluated extrinsically with two experiments. The first experiment (Section 5.2.1.) evaluates our corpus before the high-precision refinement, and is focused on DEFIE (Delli Bovi et al., 2015), an Open Information Extraction (OIE) system that works on textual definitions. In its original implementation DEFIE uses Babelfy to disambiguate definitions one-by-one before extracting relation instances. We modified that implementation and used the glosses disambiguated with our approach as input for the system, and we compared the extracted information with the information obtained by the original implementation. The second experiment (Section 5.2.2.), instead, evaluates our refined high-precision corpus, and focuses on the semantic representations of NASARI (Section 3.2.). These representations were constructed based on the BabelNet semantic network. We reimplemented NASARI using the same network enriched with the high-precision disambiguated glosses and compared these with the original glosses in the sense clustering task.

5.2.1. Open Information Extraction

In this experiment we investigated the impact of our disambiguation approach on the definitional corpus used as input for the pipeline of DEFIE. The original OIE pipeline of the system takes as input an unstructured corpus of textual definitions, which are then preprocessed one-by-one to extract syntactic dependencies and disambiguate word

	# Glosses	# Triples	# Relations
DEFIE + glosses	**150**	**340**	**184**
DEFIE	146	318	171

Table 5: Extractions of DEFIE on the evaluation sample.

	Relation	Relation Instances
DEFIE + glosses	**0.872**	**0.780**
DEFIE	0.865	0.770

Table 6: Precision of DEFIE on the evaluation sample.

senses and entity mentions. After this preprocessing stage, the algorithm constructs a syntactic-semantic graph representation for each definition, from which subject-verb-object triples (relation instances) are eventually extracted. As highlighted in Section 3.1., poor context of particularly short definitions may introduce disambiguation errors in the preprocessing stage, which then tend to propagate and reflect on both relations and relation instances. To assess the quality of our disambiguation methodology as compared to a standard approach, we modified the implementation of DEFIE to consider our disambiguated instances instead of executing the original disambiguation step, and then we evaluated the results obtained at the end of the pipeline in terms of quality of relation and relation instances.

Experimental setup. We first selected a random sample of 150 textual definitions from our disambiguated corpus (Section 4.). We generated a baseline for the experiment by discarding all disambiguated instances from the sample, and treating the sample itself as an unstructured text of textual definitions which we used as input for DEFIE, letting the original pipeline of the system carry out the disambiguation step. Then we carried out the same procedure using, instead, the modified implementation for which our disambiguated instances are taken into account. In both cases, we ran the extraction algorithm of DEFIE and evaluated the output in terms of both relations and relation instances. Following Delli Bovi et al. (2015), we relied on two human judges and performed the same evaluation procedure described therein over the set of distinct relations extracted from the sample, as well as the set of extracted relation instances.

Results. Results reported in Tables 5 and 6 show a slight but consistent improvement resulting from our disambiguated glosses over both the number of extracted relations and triples and over the number of glosses with at least one extraction (Table 5), as well as over the estimated precision of such extractions (Table 6). Context-rich disambiguation of glosses across resources and languages enabled the extraction of 6.5% additional instances from the sample (2.26 extractions on the average from each definition) and, at the same time, increased the estimated precision of relation and relation instances over the sample by ∼1%.

5.2.2. Sense Clustering

This experiment focuses on the sense clustering task. Knowledge resources such as Wikipedia or WordNet suffer from the high granularity of their sense inventories. A meaningful cluster of senses within these sense inventories would help boost the performance in different applications (Hovy et al., 2013). In this section we will explain how to deal with this issue in Wikipedia.

We integrate the high-precision version of the network as enrichment of the BabelNet semantic network in order to improve the results of the state-of-the-art system based on NASARI lexical vectors (more details of NASARI in Section 3.2.). NASARI uses Wikipedia ingoing links and the BabelNet taxonomy in the process of obtaining contextual information for a given concept. We simply enrich the BabelNet taxonomy with the high-precision disambiguated glosses (see Section 3.2.) of the target language. The high-precision disambiguated glosses are synsets that are highly semantically connected with the definiendum, which makes them particularly suitable for enriching a semantic network. The rest of the default NASARI lexical pipeline for obtaining semantic representations (lexical specificity applied to the contextual information) remains unchanged. By integrating the high-precision disambiguated glosses into the NASARI pipeline, we obtain a new set of vector representations for BabelNet synsets, increasing its initial coverage (4.4M synsets covered by the default NASARI compared to 4.6M synsets covered by NASARI enriched with our disambiguated glosses).

Experimental setup. We used the two sense clustering datasets created by Dandala et al. (2013). The task in these datasets consists of, given a pair of Wikipedia articles, to decide whether they should be merged into a single cluster or not. The first dataset (*500-pair* henceforth) contains 500 pairs of Wikipedia articles, while the second dataset (*SemEval*) consists of 925 pairs coming from a set of highly ambiguous words taken from disambiguation tasks of SemEval workshops. We follow the original setting of (Camacho-Collados et al., 2015a) and only cluster a pair of Wikipedia articles if their similarity, calculated by using the square-rooted Weighted Overlap comparison measure (Pilehvar et al., 2013), surpasses 0.5 (i.e. the middle point in the Weighted Overlap similarity scale).

Results. Table 7 shows the results of different systems in the sense clustering task. As a naive baseline we include a system which clusters all pairs. For comparison we also include the Support Vector Machine classifier of Dandala et al. (2013) exploiting information of Wikipedia in four different languages (*Dandala-multilingual*). Finally, we report the results of the default NASARI English lexical vectors (*NASARI*[13]) and the NASARI-based vectors obtained from the BabelNet semantic network enriched with our high-precision disambiguated glosses (*NASARI+glosses*). As we can see from Table 7, the enrichment produced by our glosses proved to be highly beneficial, significantly improving on the original results obtained by NASARI. Moreover, NASARI+glosses obtains the best performance overall, outperforming Dandala-multilingual in terms of accuracy in both datasets.

---

[13]Downloaded from `http://lcl.uniroma1.it/nasari/`

1705

```
<definition resource="WN" id="00166552n">
   <text>Interchanging the positions of the king and a rook</text>
      <annotations>
         <annotation source="MCS" anchor="Interchanging" bfScore="--" coherenceScore="--">bn:00087802v</annotation>
         <annotation source="MCS" anchor="positions" bfScore="--" coherenceScore="--">bn:00062704n</annotation>
         <annotation source="MCS" anchor="king" bfScore="--" coherenceScore="--">bn:00049141n</annotation>
         <annotation source="BABELFY" anchor="rook" bfScore="0.8557" coherenceScore="0.4608">bn:00016571n</annotation>
      </annotations>
</definition>
```

Figure 1: Sample XML output for the definition of *castling* in WordNet from the complete disambiguated corpus.

```
<definition resource="WN" id="00166552n">
   <text>Interchanging the positions of the king and a rook</text>
      <annotations>
         <annotation source="NASARI" anchor="king" bfScore="--" coherenceScore="--" nasariScore="0.8644">bn:00049147n</annotation>
         <annotation source="BABELFY" anchor="rook" bfScore="0.8557" coherenceScore="0.4608" nasariScore="0.8641">bn:00016571n</annotation>
      </annotations>
</definition>
```

Figure 2: Sample XML output for the definition of *castling* in WordNet from the high-precision disambiguated corpus.

	500-pair		SemEval	
	Acc.	F1	Acc.	F1
NASARI+Glosses	**86.0**	**74.8**	**88.1**	**64.7**
NASARI	81.6	65.4	85.7	57.4
Dandala-multilingual	84.4	-	85.5	-
Baseline	28.6	44.5	17.5	29.8

Table 7: Accuracy (Acc.) and F-Measure (F1) percentages of different systems on the Wikipedia sense clustering datasets.

- `anchor`: this corresponds to the exact surface form match found within the definition.

- `bfScore`: this corresponds to the Babelfy score.

- `coherenceScore`: this corresponds to the coherence score.

- `nasariScore`: this corresponds to the NASARI score (only for the high-precision annotations).

6. Release

The corpus of disambiguated glosses is freely available at http://lcl.uniroma1.it/disambiguated-glosses. We released both the *complete* (Section 3.1.) and the *high-precision* (Section 3.2.) versions of our corpus. The format for each of the two versions is almost identical: the corpus is first divided by resource (WordNet, Wikipedia, Wiktionary, Wikidata and OmegaWiki) and each resource is then divided by language.

The disambiguated glosses for each language and resource are stored in standard XML files. Figures 1 and 2 show a sample definition as displayed in the XML files of, respectively, the high-precision and complete version of our disambiguated corpus. Each file contains a list of `definition` tags, with their respective `id`[14] as attribute. Then, each definition tag is composed by the original definition as plain text and annotations. The `annotation` tag refers to the sense-annotations provided as a result of our disambiguation process. Each annotation includes its disambiguated BabelNet id and has four (or five) attributes (see Section 3. for more details about the attributes):

- `source`: this indicates whether the disambiguation has been performed by Babelfy, the Most Common Sense ("MCS") heuristic (only in the complete version of the corpus) or NASARI (only in the high-precision version of the corpus).

7. Conclusion

In this paper we presented a large-scale multilingual corpus of disambiguated glosses. Disambiguation was performed by exploiting cross-resource and cross-language complementarities of textual definitions. By leveraging the structure of a wide-coverage semantic network and sense inventory like BabelNet, we obtained a fully disambiguated corpus of textual definitions coming from multiple sources and multiple languages which, to the best of our knowledge, constitutes the largest available corpus of its kind. Additionally, we refined our sense annotations by integrating a module based on semantic similarity into our disambiguation pipeline, in order to identify a subset of high-precision disambiguated instances across the definitions. This refined version of the corpus has a great potential in high-precision low-coverage applications, where having a disambiguation error as low as possible is the first requirement. Since the disambiguated instances in this version of the corpus are directly connected to the definiendum, this high-precision disambiguated corpus may also be used to enrich a semantic network, or even used as a semantic network on its own. We evaluated our corpus intrinsically on three different languages, showing that our system outperforms previous approaches and a standard state-of-the-art disambiguation system in terms of coverage, precision and recall. We also carried out an extrinsic evaluation that shows some applications of our resource: we integrated the complete and high-precision versions of our corpus into the pipeline of both an Open Information Extraction system and a Sense Clustering system, improving on their original results and obtaining state-of-the-art figures in both tasks.

---

[14] Identifiers depend on the resource, e.g. offsets in WordNet and page titles in Wikipedia.

Acknowledgments

8. Bibliographical References

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of EACL*, pages 33–41.

Banerjee, S. and Pedersen, T. (2002). An adapted Lesk algorithm for Word Sense Disambiguation using WordNet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'02, pages 136–145, Mexico City, Mexico.

Benedictis, F. D., Faralli, S., and Navigli, R. (2013). GlossBoot: Bootstrapping Multilingual Domain Glossaries from the Web. In *Proceedings of ACL*, pages 528–538.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2015a). NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, pages 567–577.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2015b). A unified multilingual semantic representation of concepts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 741–751, Beijing, China, July.

Chen, X., Liu, Z., and Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP*, pages 1025–1035.

Chen, T., Xu, R., He, Y., and Wang, X. (2015). Improving distributed representation of word sense via wordnet gloss composition and context clustering. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers).

Dalvi, B., Minkov, E., Talukdar, P. P., and Cohen, W. W. (2015). Automatic Gloss Finding for a Knowledge Base using Ontological Constraints. In *Proceedings of WSDM*, pages 369–378.

Dandala, B., Hokamp, C., Mihalcea, R., and Bunescu, R. C. (2013). Sense clustering using Wikipedia. In *Proceedings of Recent Advances in Natural Language Processing*, pages 164–171, Hissar, Bulgaria.

Delli Bovi, C., Telesca, L., and Navigli, R. (2015). Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis. *Transactions of the Association for Computational Linguistics (TACL)*, 3.

Espinosa-Anke, L. and Saggion, H. (2014). Applying Dependency Relations to Definition Extraction. *Natural Language Processing and Information Systems*, 8455:63–74.

Espinosa-Anke, L., Saggion, H., Ronzano, F., and Navigli, R. (2016). ExTaSem! Extending, Taxonomizing and Semantifying Domain Terminologies. In *Proceedings of the 30th Conference on Artificial Intelligence (AAAI'16)*.

Fernandez-Ordonez, E., Mihalcea, R., and Hassan, S. (2012). Unsupervised word sense disambiguation with multilingual representations. In *LREC*, pages 847–851.

Flati, T., Vannella, D., Pasini, T., and Navigli, R. (2014). Two is bigger (and better) than one: the Wikipedia Bitaxonomy project. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 945–955, Baltimore, USA.

Franco-Salvador, M., Rosso, P., and Montes-y Gómez, M. (2016). A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*.

González, A., Rigau, G., and Castillo, M. (2012). A graph-based method to improve Wordnet domains. In *Proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pages 17–28, New Delhi, India.

Hill, F., Cho, K., Korhonen, A., and Bengio, Y. (2015). Learning to understand phrases by embedding the dictionary. *arXiv preprint arXiv:1504.00548*.

Hovy, E. H., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Khan, M. F., Khan, A., and Khan, K. (2013). Efficient word sense disambiguation technique for sentence level sentiment classification of online reviews. *Science International (Lahore)*, 25:937–943.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual Conference on Systems Documentation,* Toronto, Ontario, Canada, pages 24–26.

Litkowski, K. C. (2004). Senseval-3 task: Word-sense disambiguation of wordnet glosses. In *In Proc. of SENSEVAL-3 Workshop on Sense Evaluation, in the 42th Annual Meeting of the Association for Computational Linguistics (ACL 2004*. Citeseer.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. (1993). A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308.

Moldovan, D. and Novischi, A. (2004). Word sense disambiguation of wordnet glosses. *Computer Speech & Language*, 18(3):301–317.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Navigli, R. and Velardi, P. (2005). Structural Semantic Interconnections: a knowledge-based approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1088.

Navigli, R. and Velardi, P. (2010). Learning Word-Class Lattices for definition and hypernym extraction. In *Proceedings of ACL 2010*, pages 1318–1327, Uppsala, Sweden.

Navigli, R. (2009). Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Novischi, A. (2002). Accurate semantic annotations via pattern matching. In *FLAIRS Conference*, pages 375–379.

Pilehvar, M. T., Jurgens, D., and Navigli, R. (2013). Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of ACL*, pages 1341–1351.

Richardson, S. D., Dolan, W. B., and Vanderwende, L. (1998). MindNet: Acquiring and Structuring Semantic Information from Text. In *Proceedings of ACL*, pages 1098–1102.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Velardi, P., Faralli, S., and Navigli, R. (2013). OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3):665–707.

9. Language Resource References

Miller, G. A. and Beckwith, R.T. and Fellbaum, Christiane D. and Gross, D. and Miller, K. (1990). *WordNet: an Online Lexical Database*.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.